Learning a Model of Speaker Head Nods using Gesture Corpora

Jina Lee
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
jlee@ict.usc.edu

During face-to-face conversation, the head is constantly in motion, especially during speaking turns [2]. These movements are not random; research has identified a number of important functions served by head movements [7] [5] [3] [4]. Head movements provide a range of information in addition to the verbal channel such as nods to show our agreement or shakes to express disbelief.

The goal of our work is to build a domain-independent model of speaker's head movements and use the model to generate head movements for virtual agents. To use the model for interactive virtual agents, it needs to operate in real-time. For this reason, we focus on features that are readily available at the time head movements are generated. In addition, we plan to make the model portable to other systems by using features such as part of speech tags that are easily obtainable even when using different language tools.

In this paper, we present a data-driven, automated approach to generate speaker nonverbal behavior, which we demonstrate and evaluate by learning when head nods should occur. Specifically, the approach uses a machine-learning technique (i.e. learning a hidden Markov model [8]) to create a head nod model from annotated corpora of face-to-face human interaction, relying on the linguistic features of the surface text. Figure 1 illustrates the overview of the procedures to learn the model. Once the patterns of when people nod are learned, then it can be used to generate head nods for virtual agents by encoding a new sample with the factors used for learning and feeding it to the model to obtain the most likely head movement.

1. HEAD NOD PREDICTION MODEL

1.1 Gesture Corpus

For this work, we used the AMI Meeting Corpus [1]. It is a set of multi-modal meeting records, which includes 100 meeting hours. The corpus includes annotations of meeting context such as participant IDs and topic segmentations as well as annotations on each participant's transcript and movements. Annotations of each meeting are structured in an XML format and are cross-referenced through meeting IDs, participant IDs, and time reference. For this work, we used the recordings of 17 meetings, each consisted of three to four participants, which adds up to be around eight hours of meeting annotation.

1.2 Data Alignment and Feature Selection

Among all the annotations included in the corpus, we used the transcript of each speaker, the dialog acts of each utterance, and the type of head movements observed while the utterance was spoken. The head types annotated in the corpus are: nod, shake, nodshake, other, and none. We also obtained the part of speech tags and phrase boundaries (e.g. verb phrases and noun phrases) by sending the utterances through a natural language parser. In addition, we combined the features from our previous work in Nonverbal Behavior Generator (NVBG) [6], which is a rule-based system that analyzes the agent's cognitive processing and the syntactic and semantic structure of the surface text to generates nonverbal behaviors for virtual humans. We looked for keywords that trigger the rules associated with head nods in NVBG and called those keywords *key lexical entities*. From the 17 meeting recordings we used, we collected 10,000 sentences and wrote a script to cross-reference the corresponding annotation files and aligned the features on a word level.

When training hidden Markov models, we want to keep the number of features low by eliminating uncorrelated features when given a limited number of data samples. Therefore, we reduced the number of features by counting the frequency of head nods that occurred with each feature and selected a subset of them. Based on the results of the frequency counts, the final features selected for training are:

- **Part of Speech**: Conjunction, Proper Noun, Adverb, Interjection, Remainder
- **Dialog Act**: BackChannel, Inform, Suggest, Remainder
- **Sentence Start**: y, n
- **Noun Phrase Start**: y, n
- **Verb Phrase Start**: y, n
- **Key Lexical Entities**: y, n

1.3 Training Process

To learn the head nod model, hidden Markov models (HMM) were trained. For this work, the input is a sequence of feature combinations representing each word. The sequential property of this problem led us to use HMMs to predict head nods. After aligning each word of the utterances with the selected features as described above, trigrams of these words were formed as the data set. For each trigram, the head type was determined by the majority vote method; if more than two out of three words co-occurred with a nod,

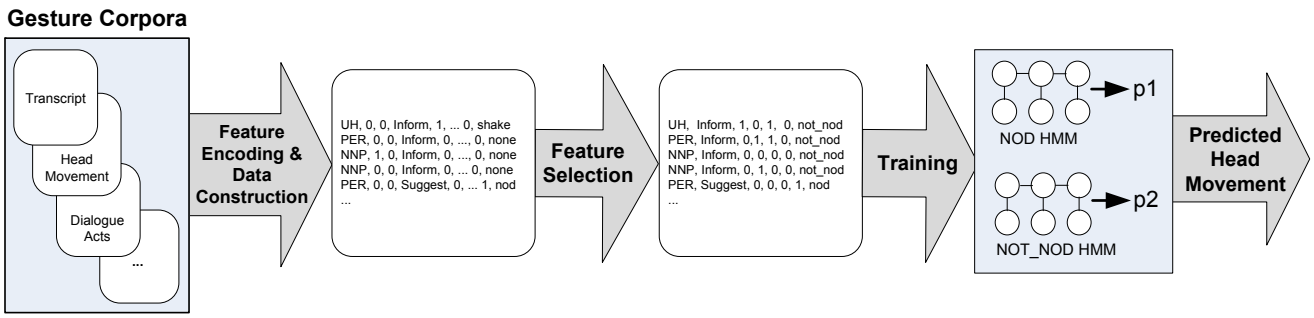

Figure 1: Overview of the head nod prediction framework. The information in the gesture corpus is encoded and aligned to construct the data set. The feature selection process chooses a subset of the features that are most correlated with head nods. Using these features, probabilistic sequential models are trained and utilized to predict whether or not a head nod should occur.

Measurement	Equation	Value
Accuracy	(tp+tn) / (tp+fp+tn+fn)	.8528
Precision	tp / (tp+fp)	.8249
Recall	tp / (tp+fn)	.8957
F-measure	2*precision*recall / (precision+recall)	.8588

Table 1: Measurements for the performance of the learned model.

the trigram was classified as a nod instance. To determine whether a trigram should be classified as a nod, we trained two HMMs, a 'NOD HMM' and a 'NOT_NOD HMM,' and fed the same trigram into both models and compared the results of each model.

To train a 'NOD HMM,' we collected all the positive instances of 'nod' trigrams from the entire set of trigrams. Then, we left out 20% of the 'nod' trigrams as a test set, which is used in the final evaluation step, and used the remaining 80% of the data for training. To train the 'NOD HMM,', we performed a standard 10-fold cross-over validation. Similarly, we repeated these steps to train the 'NOT_NOD HMM.' Finally, we ran the test set (20% of the entire data left out) through the 'NOD HMM' and 'NOT_NOD HMM' and classified each sample to have the head movement of whichever model produced a higher probability.

1.4 Results and Conclusion

To measure the performance of our learned model, we computed the accuracy, precision, recall, and F-measure of the learned model. Table 1 summarizes the results with the equations used for computing the measurements. The results show that the model can predict head nods with high precision, recall, and accuracy rate even without a rich markup of the surface text (i.e. only using the syntactic/semantic structure of the utterance and dialog act).

This work could be extended in several ways. Currently we are working on detecting the emotional state from each utterance and adding this into the feature set to investigate whether emotional data improves the learning. Further analysis of the linguistic structure may also be performed using additional language tools to extract features such as emphasis points and contrast points. We can also extend the work by learning the patterns of different head movements other than nods. Finally, we plan to conduct evaluations with human subjects to investigate if the head movements generated by the model are perceived to be natural.

Acknowledgments

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

2. REFERENCES

[1] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[2] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983.

[3] U. Hadar, T. J. Steiner, and F. C. Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[4] D. Heylen. Challenges ahead: Head movements and other social acts in conversations. In *AISB 2005, Social Presence Cues Symposium*, 2005.

[5] A. Kendon. Some uses of the head shake. *Gesture*, 2:147–182(36), 2002.

[6] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In *In Proceedings of the 6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA*, pages 243–255. Springer, 2006.

[7] E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878(24), June 2000.

[8] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.